Optimizing Information Exchange in Cooperative Multi-agent Systems

Claudia V. Goldman
Department of Computer Science
University of Massachusetts
Amherst, MA 01003
clag@cs.umass.edu

Shlomo Zilberstein
Department of Computer Science
University of Massachusetts
Amherst, MA 01003
shlomo@cs.umass.edu

ABSTRACT

Decentralized control of a cooperative multi-agent system is the problem faced by multiple decision-makers that share a common set of objectives. The decision-makers may be robots placed at separate geographical locations or computational processes distributed in an information space. It may be impossible or undesirable for these decision-makers to share all their knowledge all the time. Furthermore, exchanging information may incur a cost associated with the required bandwidth or with the risk of revealing it to competing agents. Assuming that communication may not be reliable adds another dimension of complexity to the problem.

This paper develops a decision-theoretic solution to this problem, treating both standard actions and communication as explicit choices that the decision maker must consider. The goal is to derive both action policies and communication policies that together optimize a global value function. We present an analytical model to evaluate the trade-off between the cost of communication and the value of the information received. Finally, to address the complexity of this hard optimization problem, we develop a practical approximation technique based on myopic meta-level control of communication.

Categories and Subject Descriptors

I.2.11 [**Artificial Intelligence**]: Distributed Artificial Intelligence, Control Methods; G.3 [**Mathematics of Computing**]: Probability and Statistics — Markov processes, Stochastic processes.

General Terms

Algorithms, Design, Performance, Theory.

Keywords

Decentralized Control, Communication, Robot Teams.

AAMAS'03, July 14–18, 2003, Melbourne, Australia.

1. INTRODUCTION

Conceptually, it is possible to separate the planning involved in a multi-agent system into two stages: 1) the planning that occurs before execution starts (sometimes referred to as off-line) and 2) the planning that occurs during execution time (sometimes referred to as on-line). Each of those cases can be done in a centralized or decentralized manner leading to four possible approaches.

The first approach considers centralized MAS, where both the off-line planning stage and the on-line stage are controlled by a central entity, or by all the agents in the system, who are assumed to have full observability (e.g., MMDPs [2]). In the DAI literature, *cooperative* systems were usually associated with this approach. The other well studied approach to cooperative MAS is located at the other extreme, where planning and execution are done in a completely distributed manner. These are distributed cooperative MAS. Each agent decides on its own actions based on its local knowledge and any other information it may obtain from the other agents (e.g., AMMs [7], [13] where observability of the other agents' behaviors may be obtained at execution time from feedback estimators, and [14] where individual agents learn independent policies). This class of systems has been mostly studied in non-cooperative scenarios [12, 11] where self-interested agents could not have been assumed to synchronize their actions before they start operating.

Our work focuses on the third approach: decentralized cooperative MAS. Agents in most cooperative MAS are limited by not being able to fully communicate during execution (due to the distributed aspect of the MAS), but due to the cooperative nature of the MAS, in many situations these constraints do not apply to the pre-execution stage. Thus, cooperative agents are able to share information at the off-line planning stage as if they were centrally controlled. But unlike the first approach, these agents will be acting in real-time in a decentralized manner. The agents take this into account while planning off-line. If planning is possible on-line, it will also be done in a decentralized manner.[1] For example, robots that are cooperatively programmed may be deployed to perform independent missions, although they may need to exchange local information from time to time

[1]The fourth approach that is decentralized before execution and centralized during execution has received little attention. This is also an interesting case in which agents work in a decentralized manner on a plan and then they implement it in a centralized manner assuming full-observability during execution.

to coordinate more efficiently and to achieve better performance as a group.

Claus and Boutilier [3] studied a simple case of decentralized control where agents have full observability of the other agent's actions during the off-line planning stage. The solution presented in their example includes a joint policy of a single action for each agent to be followed in a stateless environment. The agents learn which equilibrium to play. In our model, partial observability is assumed and the scenarios studied are more complex and include states. There are no solutions to the general decentralized control problem with communication, which is addressed in this paper.

In all the above cases, when agents can communicate, a certain fixed communication language is assumed to be known to the agents. Recent work on language evolution [17, 6] produced algorithms for agents to develop their own languages of communication. The agents in Wang and Gasser's model learn a mutual concept and they do not consider the problem of learning the language in the framework of planning their control actions. Agents in [6] learn ontologies related to their actions, but the language problem is studied separately from the control problem.

Real life situations involving more than one decision-maker are frequently characterized by time-constrained operation and uncertainty. Examples include autonomous exploration or monitoring of a target environment, rejoining units acting in unknown territories, factory operation where machines may act independently with some need for planning and coordination, and rescue operations where agents decide dynamically what areas to cover for searching and mapping.

While much progress has been made in the area of reasoning under uncertainty by a single decision-maker, there are no adequate solutions for the decentralized version of the problem. This paper focuses on cooperative multi-agent systems where the agents share a common goal. In order to achieve the global goal of the system (e.g., load-balancing of a system, achieve greater efficiency in resource allocation, or achieve physical coordination of robots) the agents may need to communicate to synchronize their information. However, it may be impossible or undesirable for these agents to share all their knowledge all the time. Exchanging information may incur a cost associated with the required bandwidth or with the risk of revealing it to competing agents. Communication may also be unreliable. This paper focuses on agents that may need some of this information to get synchronized from time to time, but they cannot assume that communication is free and information can be exchanged at each moment. Most current systems rely on ad-hoc heuristics (e.g., [8]), or they rely on the assumption that knowledge can be shared constantly (e.g., [4, 10, 5]). Xuan et al.[18] address the problem of combining communication acts into the decision problem of a group of cooperative agents. Their framework is similar to ours but their approach is heuristic. We are interested in the most comprehensive case where cooperative agents must determine which messages they should transmit, and when they should transmit them, assuming communication incurs a cost.

Section 2 presents our formal approach to decentralized control with communication and discusses the relevant aspects of choosing different types of messages. A communication language together with communication acts are added to the decentralized model for partially observable Markov Decision Processes[1]. The larger goal of this research is to address the design of languages of communication, including their semantics and their complexity (e.g., agents may transmit only signals or state-observations, or they may communicate at higher levels of reasoning such as sending strategies of behaviors).

The COM-MTDP model [15] offers a framework that considers the uncertainties and costs in real-world scenarios, addressing some of the deficiencies of BDI systems. The authors compare complexity results when either free communication, no communication or general communication is assumed. While the model accounts for the cost of communication, it does not consider the different cost models that we examine in this work. We give a formal statement of the problem when a certain language and semantics are assumed, including the full formulation of the value of an optimal policy of action and of communication. Pynadath and Tambe applied a single case of communication, which allows an agent to send a single message indicating that a certain goal has been achieved. Our work studies a more general problem: the agents optimize the timing and frequency of communication, and are allowed to communicate more than once. We are interested in problems where agents may act independently to achieve their own tasks, but may need to synchronize their knowledge from time to time to coordinate in more efficient ways. This includes each agent deciding *when* and *what* to communicate.

Section 3 presents a practical and feasible approach to the communication control problem in decentralized cooperative MAS aimed at finding approximate solutions based on a myopic approximation. The greedy approach optimizes the choice of a single message ignoring future messages and iterates on this policy to find an approximate solution to the general problem. The resulting joint value of the actions and communication policies with multiple transmissions of messages depends on the cost incurred by sending a message, and the cost of taking a control action. The empirical results show the range in utility values that can be attained at worst when no communication is assumed and a single fixed goal is given and at best when constant and free communication is assumed. Within this range, we describe the performance of our approximation algorithm compared to a heuristic case based on sub-goal communication independent of its cost. For larger costs of communication the approximation attained by iterating on an optimal single communication may lead to suboptimal solutions, for which some deterministic policy of communication may be more beneficial if the right parameters are known. In case that the optimal parameters for the heuristic are not known, the greedy approach yields higher joint utility values.

Optimizing both the control actions and the communication policy is a very complex problem as supported by [1] and [15]. The framework presented in Section 2 also emphasizes the amount of information that is required from an agent to compute analytically a joint policy. However, it remains to be verified whether other attempts may find a tractable solution to the decentralized control with communication problem. This paper is one of the few formal studies done to tackle this hard problem. It presents a promising direction based on greedy meta-level control of communication that has also proved useful in meta-level control of computation (e.g., [16]) and information gathering [9], in which non-myopic control is extremely difficult.

2. THE THEORETICAL FRAMEWORK

We present a formal model for decentralized control that is based on an extension of the decentralized partially-observable Markov Decision Process. Within the model, cooperative agents are represented by finite state controllers, whose actions control the process. We focus on a special type of decentralized partially-observable Markov Decision Process with communication, *Dec_POMDP_Com*, defined as follows:[2] $M^{com} =< S, A_1, A_2, \Sigma, C_\Sigma, P, R, \Omega_1, \Omega_2, O, T >$ where:

- S is a finite set of states. s_0 is the initial state of the system. Each state $s = (s_1, s_2)$ where $s_i \in S_i, i = \{1, 2\}$ are the local states of the corresponding agents.
- A_1 and A_2 are finite sets of control actions. a_i denotes the action performed by agent i.
- Σ is the alphabet of messages. $\sigma_i \in \Sigma$ denotes an atomic message sent by agent i. $\overline{\sigma_i}$ is a sequence of atomic messages sent by agent i. A special message that belongs to Σ is the null message which is denoted by ϵ_σ. This message is sent by an agent that does not want to transmit anything to the other agents. The agents do not incur any cost in sending a null message.
- C_Σ is the cost of transmitting an atomic message. $C_\Sigma(\epsilon_\sigma) = 0$, $C_\Sigma : \Sigma \rightarrow \Re$.
- P is the transition probability function. $P(s, a_1, a_2, s')$ is the probability of moving from state s to state s' when agents 1 and 2 perform actions a_1 and a_2.
- R is the reward function. $R(s, a_1, \sigma_1, a_2, \sigma_2, s')$ represents the reward obtained by the system as a whole, when agent 1 executes action a_1 and sends message σ_1, and agent 2 executes action a_2 and sends message σ_2 in state s resulting in a transition to state s'.
- Ω_1 and Ω_2 are finite sets of observations.
- O is the observation function. $O(s, a_1, a_2, s', o_1, o_2)$ is the probability of observing o_1 and o_2 (respectively by the two agents) when in state s agent 1 takes action a_1 and agent 2 takes action a_2, resulting is state s'.
- T is a positive integer representing the horizon.

A *Dec_POMDP_Com* is *jointly fully-observable* if there exists a mapping $J : \Omega_1 \times \Omega_2 \rightarrow S$ such that whenever $O(s, a_1, a_2, s', o_1, o_2)$ is non-zero then $J(o_1, o_2) = s'$.

In addition to this notion known for general *Dec_POMDP*s, we consider another property of our framework due to the communication involved. A *Dec_POMDP_Com* is *jointly synchronized* if both agents have the same knowledge about the global state, and none of the agents separately has more knowledge than this. This knowledge is not necessarily the global state itself. Joint full observability is a special case of joint synchronization. Agents may be jointly synchronized if they both know with certainty some features of the global state. Note that in our model, communication is the *only* means of achieving synchronization. There is no other way in which a single agent will perform an action and will be able to get the other agent's observation.

Agents may need histories of observations to become synchronized. The information corresponding to the union of the histories of observations give the agents information about their current global state.

We describe the interaction among the agents as a process in which agents perform an action, then they observe their environment, and then send a message that is instantaneously received by the other agent.[3] Then, we can define the local policies of the controlling agents as well as the resulting joint policy whose value we are interested in optimizing. A *local policy* δ is composed of two policies, δ^A that determines the actions of the agent, and δ^Σ that states the communication policy.

DEFINITION 1. *A local policy for action for agent i, δ_i^A is a mapping from local histories of observations $\overline{o_i} = o_{i_1}, \ldots, o_{i_t}$ over Ω_i and histories of messages $\overline{\sigma_j} = \sigma_{j_1}, \ldots, \sigma_{j_t}$ received $(j \neq i)$ since the last time the agents were synchronized to actions in A_i.*

$$\delta_i^A : S \times \Omega^* \times \Sigma^* \rightarrow A_i$$

DEFINITION 2. *A local policy for communication for agent i, δ_i^Σ is a mapping from local histories of observations $\overline{o_i} = o_{i_1}, \ldots, o_{i_t}$ and o, the last observation perceived after performing the last local action, over Ω_i and histories of messages $\overline{\sigma_j} = \sigma_{j_1}, \ldots, \sigma_{j_t}$ received $(j \neq i)$ since the last time the agents were synchronized to messages in Σ.*

$$\delta_i^\Sigma : S \times \Omega^* o \times \Sigma^* \rightarrow \Sigma$$

DEFINITION 3. *A joint policy $\delta =< \delta_1, \delta_2 >$ is defined to be a pair of local policies, one for each agent, where each δ_i is composed of the communication and the action policy for agent i.*

2.1 Characteristics of the Model

The model studied in this paper is further characterized by including *decomposable actions*, i.e., both the transitions and the observations are independent: $P(s_1'|s_1, a_1) = P(s_1'|s_1, s_2, a_1, a_2)$ and $O(o_1|s, a_1, a_2, s') = O(o_1|s_1, a_1, s_1')$ where s_i and s_2 correspond to each agent's partial view of the global state, meaning that the agents do not affect the distribution of the outcome states of each other nor their observations (notice that this assumption does not mean that the problem the agents are solving is decomposable), *reliable* messages (i.e., any message σ_i sent by agent i is reliably received by the other agent). We assume that both agents remember the sequences of atomic messages received (i.e., both agents store histories of messages). This may be beneficial when the agents do not have complete joint observability even when they synchronize. Agents send their messages in a broadcast way, all of the agents receive all of the messages sent. Finally, communication is the only means of achieving joint full observability and joint synchronizability. The model presented so far can be specialized along three different dimensions:

- The language Σ — What are the messages that the agents can transmit to each other?
- The types of communication among the agents — How is the reward function of the Dec_POMDP_Com affected by the cost model of the communication?
- The communication protocols — The general control problem optimizes over all possible protocols, but this framework allows also the analysis of certain families of protocols.

[2]This paper is restricted to a case with two agents. The model can be extended to any number of agents.

[3]When agents exchange information there is a question whether information is obtained instantaneously or there are delays. For simplicity of exposition we assume no delays in the system.

2.2 The communication language Σ

The semantics of a message can be attached explicitly by the agents' designer. In addition to this explicit semantics, there exists an implicit semantics for each message, given by the context in which this message was received. The context of a message sent by agent j is a pair consisting of 1) the last synchronized state when the agents exchanged their information, and 2) the sequence of observations observed by agent i when agent j sent this message to him. A local policy defines implicitly the triplets of synchronized states, observation sequences and messages, and therefore the table of instantiations of the δ mapping gives us all the possible meanings for the language Σ. In this paper, we assume that agents communicate messages of the same length and that the agents understand the messages received. These restrictions may be relaxed when solving the general problem of decentralized control with communication. Two cases can be distinguished when studying a Dec_POMDP_Com:

1. The general case — Σ is the most general language, i.e., an explicit semantics was not given a-priori. The decentralized control problem with unrestricted-semantics communication can be defined as follows:

PROBLEM 1. *Find a joint policy that maximizes the expected total reward over the finite horizon. Solving for this policy embeds* the optimal meanings *of the messages chosen to be communicated.*

Since semantics are given by the context in which the messages are sent, the empty message may have more than one meaning. However, this null message does not incur a cost when transmitting it. This fact raises interesting research questions with respect to agents that can optimize their actions also when they do not receive any message. Agent i decides not to send a message considering that agent j may *understand* i's state even though j does not receive any message from i. Notice also that even though no fixed semantics are assumed, the context of a message can serve as a signal for an agent to take some action.

2. Σ has a fixed semantics — Designers that set a fixed semantics to a given language Σ allow the agents to achieve performance of certain actions by exchanging messages with known meanings. More complex studies of these languages may allow for agents to reason at a meta-level of these given semantics. That is, agents may take an action following a message they have received, and move eventually to a new state. Observing this new state, may lead the agent to take an additional action that was not the original aim of the message sent, but it is the result of an effect the sender of the message had on the resulting state. One example of a language with fixed semantics is the language of observations (i.e., $\Sigma_i = \Omega_i$), where the agents communicate their observations. The decentralized control problem with fixed-semantics communication can be defined as follows:

PROBLEM 2. *Find a joint policy that maximizes the expected total reward over the finite horizon that consists of the policy of actions and policy of communications* given the semantics *fixed for the language of communication.*

2.2.1 Types of Messages

In this paper, we focus on *informative messages*. Informative messages affect the decision of the hearer when he chooses its next action. This is expressed in the δ_i^A function. δ_i^A is a function of the last synchronized state, the last observation o and all the r messages received so far from agent j. These last messages affect directly the decision of agent i. On the other hand, informative messages do not affect the outcome of the action chosen by the hearer. This results from the definition of the transition probability function, $P(s, a_1, a_2, s')$, which depends only on the actions performed by the agents, and not on the messages transmitted.

Other types of messages include the following:

1. Commitments — a message sent by an agent expressing its commitment to doing a certain action at a certain time, or its request for a commitment.

2. Reward/punishment — an encouraging or punishing signal that is sent to another agent, which can be eventually considered for learning adaptive behaviors.

3. World information — Both agents are assumed to have prior knowledge about the model of the world. But agents may be willing to exchange information regarding their policies of actions for the future, for example.

Finding optimal joint policies that consider these last kinds of messages remain for future research.

2.3 Types of Communication

Types of communication determine the flow of the information exchange and the cost of this communication. We describe three possible cost models that can be framed in the Dec_POMDP_Com model. In Section 3 the two-way communication model was implemented.

1. One-way communication occurs when agent i sends information to agent j. The information flow is unidirectional. If both agents communicate at the same state, then they will incur two costs for achieving this communication.

2. Two-way communication necessarily leads to joint exchange of messages. Both agents incur just one cost to attain the information.

3. Acknowledged communication requires a confirmation should be sent for every message that was received. This acknowledgment can be understood as an agreement on the part of the hearer agent to the message sent, and its consent to act as expected. More generally, this acknowledgment assures that the hearer has received the message sent to it when communication may be unreliable.

2.4 Communication Protocols

Communication protocols determine when the agents should communicate and what information should they transmit. Solving the general decision problem should also solve for the optimal protocol. But a simpler decision question can be answered for a given protocol as is shown in Section 3.2 for a special case in which agents exchange information when a sub-goal is achieved.

2.5 The Value of the Optimal Joint Policy

Following the model presented in Section 2, we solve for the value of a state in the *Dec_POMDP_Com*. The optimal joint policy that stipulates for each decision-maker how it should behave and when it should communicate with other agents is the policy that maximizes the value of the initial state of the *Dec_POMDP_Com*.

In order to refer to a sequence of messages sent by an agent, two auxiliary functions are defined: f_1^i are the first i messages sent by agent 1 with length equal to i. Similarly, f_2^i is defined for the messages sent by agent 2. f_1^i is a function

of 1) the state in which the last message is sent, 2) the sequence of observations seen by agent 1 (when $|\overline{o_1}| = i$, is denoted by $\overline{o_1}^i$), and 3) the sequence of messages received from agent 2. These functions can be recursively defined:

$$f_1^0 = \delta_1^\Sigma(s, \epsilon, \epsilon) \quad f_2^0 = \delta_2^\Sigma(s, \epsilon, \epsilon)$$
$$f_1^i = \delta_1^\Sigma(s, \overline{o_1}^{i-1}, f_2^{i-1}) \cdot f_1^{i-1}$$
$$f_2^i = \delta_2^\Sigma(s, \overline{o_2}^{i-1}, f_1^{i-1}) \cdot f_2^{i-1}$$

DEFINITION 4. *The probability of transitioning from a state s to a state s' following the joint policy $\delta =< \delta_1, \delta_2 >$ while agent 1 sees observation sequence $\overline{o_1}o_1$ and receives sequences of messages $\overline{\sigma_2}$, and agent 2 sees $\overline{o_2}o_2$ and receives $\overline{\sigma_1}$ of the same length, written $\overline{P_\delta}(s, \overline{o_1}o_1, \overline{\sigma_2}, \overline{o_2}o_2, \overline{\sigma_1}, s')$ can be defined recursively:*[4]

1. $\overline{P_\delta}(s, \epsilon, \epsilon, \epsilon, \epsilon, s) = 1$
2. $\overline{P_\delta}(s, \overline{o_1}o_1, \overline{\sigma_2}\sigma_2, \overline{o_2}o_2, \overline{\sigma_1}\sigma_1, s') =$

$$\sum_{q \in S} \overline{P_\delta}(s, \overline{o_1}, \overline{\sigma_2}, \overline{o_2}, \overline{\sigma_1}, q).$$
$$P(q, \delta_1^A(s, \overline{o_1}, \overline{\sigma_2}), \delta_2^A(s, \overline{o_2}, \overline{\sigma_1}), s').$$
$$O(q, \delta_1^A(s, \overline{o_1}, \overline{\sigma_2}), \delta_2^A(s, \overline{o_2}, \overline{\sigma_1}), s', o_1, o_2)$$
$$\text{such that } \delta_1^\Sigma(s, \overline{o_1}o_1, \overline{\sigma_2}) = \sigma_1 \wedge \delta_2^\Sigma(s, \overline{o_2}o_2, \overline{\sigma_1}) = \sigma_2.$$

Then, the value of a state s in the *Dec_POMDP_Com* from following a joint policy δ for T steps can be defined as follows:

DEFINITION 5. *The value $V_\delta^T(s)$ of following policy $\delta =< \delta_1, \delta_2 >$ from state s for T steps is given by:*

$$V_\delta^T(s) = \sum_{<\overline{o_1}o_1, \overline{o_2}o_2>} \sum_{q \in S} \sum_{s' \in S} \overline{P_\delta}(s, \overline{o_1}, f_2^i, \overline{o_2}, f_1^i, q).$$
$$P(q, \delta_1^A(s, \overline{o_1}, f_2^i), \delta_2^A(s, \overline{o_2}, f_1^i), s').$$
$$R(q, \delta_1^A(s, \overline{o_1}, f_2^i), \delta_1^\Sigma(s, \overline{o_1}o_1, f_2^i), \delta_2^A(s, \overline{o_2}, f_1^i), \delta_2^\Sigma(s, \overline{o_2}o_2, f_1^i), s')$$

where the observation and the message sequences are of length at most $T-1$, and both sequences of observations are of the same length i. The sequences of messages will then be of length $i+1$ because they considered the last observation resulting from the control action previous to communicating.

PROBLEM 3. *The decentralized control problem with communication is to find an optimal joint policy δ^* for action and for communication such that $\delta^* = argmax_\delta V_\delta^T(s_0)$.*

The finite horizon problem has been proved to be NEXP-complete [1] and the infinite version of the problem is known to be undecidable.

3. MYOPIC GREEDY COMMUNICATION

The complexity analysis done for decentralized control in the framework of *Dec_POMDP*s [1] covers the worst case scenarios which also include communication among the agents at a prohibitively expensive cost. In the other extreme, assuming free communication at every moment transforms the decentralized control problem into a Markov Decision Process control problem, which is known to be tractable.

The gap between these two extreme cases is worth studying for reasonable costs of communication. Real-life applications that involve the need for information exchange have usually a bounded cost that measures real bandwidth cost or it may measure the risk attached to the communication act. We study how to trade off the cost or risk of communication with its benefits. In terms of complexity, the *Dec_POMDP_Com* approach is closer to the MDP extreme rather than to the general decentralized POMDP with prohibitively high cost of communication. Therefore, we expect that our approach will indeed find tractable optimal as well as approximate solutions to the decentralized control problem when communication is explicitly modeled and feasible.

3.1 Meeting under Uncertainty Example

The first problem studied as a *Dec_POMDP_Com* involves two agents that have to meet at some location as early as possible. The environment is represented by a 2D grid with discrete locations. Both agents know each other's location at time t_0. They set the meeting point G_{t_0} at the middle of their Manhattan distance d_0. Each agent can move towards G_{t_0} following the optimal policy of action each can compute. We assume each agent moves independently in the environment. In this setting, there is uncertainty regarding the outcomes of the agents' actions, i.e., with probability P_u, an agent arrives at the desired location after having taken a move action, but with probability $1-P_u$ the agent remains at the same location. Due to this uncertainty in the agents' actions' effects, it is not clear that setting a predetermined meeting point to which the agents will optimally move is the best strategy for designing these agents. Agents may be able to meet faster if they change their meeting place after realizing their actual locations. This can be achieved by exchanging information about the locations of the agents, that otherwise are not observable.

We study the case where agents can communicate their observations (i.e., their actual locations), incurring a cost C_Σ. Both agents become synchronized if at any time t, at least one agent initiates communication. Joint-exchange of messages is assumed (see Section 2.3). The agents will set a new goal G_t that is computed as the middle of the distance between the agents revealed at time t. The utility attained by both agents in the following four different scenarios are compared:

1. No-Communication — Each agent follows its optimal policy of action without communication. The meeting point is fixed at time t_0 and cannot be changed.
2. Ideal — Assuming $C_\Sigma = 0$, and the agents communicate at each time step, this is the highest joint utility both agents can attain. Notice, though, that this is not the optimal solution we are looking for, because we do assume that communication is not free. Nevertheless, the difference in the utility obtained in these first two cases shed light on the trade-off that can be achieved by implementing non-free communication policies.
3. Communicate Sub-Goals — A heuristic solution which assumes that the agents have a notion of sub-goals. They notify each other when these sub-goals are achieved, eventually leading the agents to meet.

[4]The notation $\overline{o} = o_1, \ldots, o_t$ and $\overline{o}o$ represents the sequence $o_1, \ldots, o_t o$. Similarly, the notation for sequences of messages: $\overline{\sigma_i}\sigma$ represents the sequence $\sigma_{i_1}, \ldots, \sigma_{i_t}\sigma$.

4. Greedy Approach — Agents act myopically optimizing the choice of when to send a message, assuming no additional communication is possible. For each possible distance between the agents, a policy of communication is computed such that it stipulates when it is the best time to send that message. By iterating on this policy agents are able to communicate more than once and thus approximate the optimal solution to the decentralized control with communication problem.

3.2 Experiments

The No-Communication case consists of the agents setting a fixed goal at the middle of the distance between them at time t_0. Both agents move optimally towards this goal.[5] This problem can be solved analytically by computing the expected cost[6] $\Theta_{nc}(A, d_1, d_2)$ incurred by two agents located at distances d_1 and d_2 respectively from the goal at time t_0. $\Theta_{nc}(A, 0, 0) = 0$ and in general:[7]
$\Theta_{nc}(A, d1, d2) =$

$$\frac{1}{P_u(2-P_u)}[2R_a + P_u^2\Theta_{nc}(A, d1-1, d2-1) + P_u(1-P_u)\Theta_{nc}(A, d1-1, d2) + (1-P_u)P_u\Theta_{nc}(A, d1-, d2-1)]$$

$\Theta_{nc}(A, 0, d2) = \frac{1}{P_u}[2R_a + P_u\Theta_{nc}(A, 0, d2-1)]$
$\Theta_{nc}(A, d1, d0) = \frac{1}{P_u}[2R_a + P_u\Theta_{nc}(A, d1-1, 0)]$

In the Ideal case, a set of 1000 experiments was run in which the cost of communication was assumed to be zero. Agents communicate their locations at every instance, and update the location of the meeting place accordingly. Agents move optimally to the last synchronized meeting location.

For the third case tested (Communicate Sub-Goals) a sub-goal was defined by the cells of the grid with distance equal to $p * d/2$ from the fixed current meeting point. d expresses the Manhattan distance between the two agents, this value is accurate only when the agents synchronize their knowledge. That is at time t_0 the agents determine the first sub-goal as the area bounded by a radius of $p * d_0/2$ and which center is located at $d_0/2$ from each one of the agents. Each time t that the agents synchronize their information through communication, a new sub-goal is determined at $p * d_t/2$. p is a parameter of the problem that determines the radius of the circle that will be considered a sub-goal and therefore induces the communication strategy. Figure 1 shows how new sub-goals are set when the agents transmit their actual location once they reached a sub-goal area. The meeting point is dynamically set to be at the middle of the last synchronized Manhattan distance between the agents.

Experiments were run for the Communicate Sub-Goals case for different uncertainty values and different costs of communication. These results show that agents can obtain higher utility by adjusting the meeting point dynamically rather than having set one fixed meeting point. Agents can synchronize their knowledge and thus they can set a new

[5]Each agent has an associated MDP in the given model.
[6]Cost and utility are used interchangeably as appropriate meaning cost is minimized and utility is maximized.
[7]The cost of taking one control action is R_a. The agents aiming at meeting, are trying to minimize the time to meet. Therefore, the agents incur a cost of $R = 2R_a$ as long as they have not actually met (i.e., even though only one of the agents might have reached the goal first).

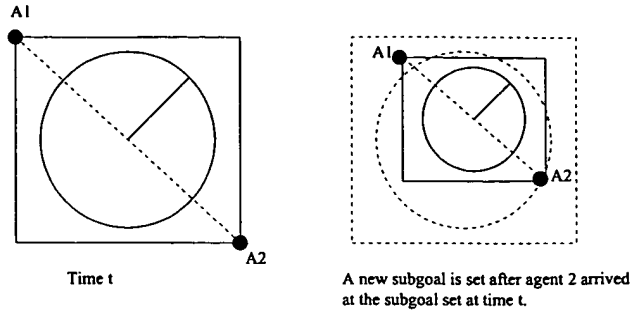

Figure 1: Goal decomposition into sub-goal areas.

meeting location instead of acting as two independent MDPs that do not communicate and move towards a fixed meeting point (see Figure 2). Notice that there are still cases for certain values of p where the joint utility of the agents is actually smaller than the joint utility achieved in the No Communication case (2 MDPs). This points out the need to empirically tune the parameters needed in the implemented heuristic, as opposed to a formal approach to approximate the solution to the problem as is shown in the Greedy case.

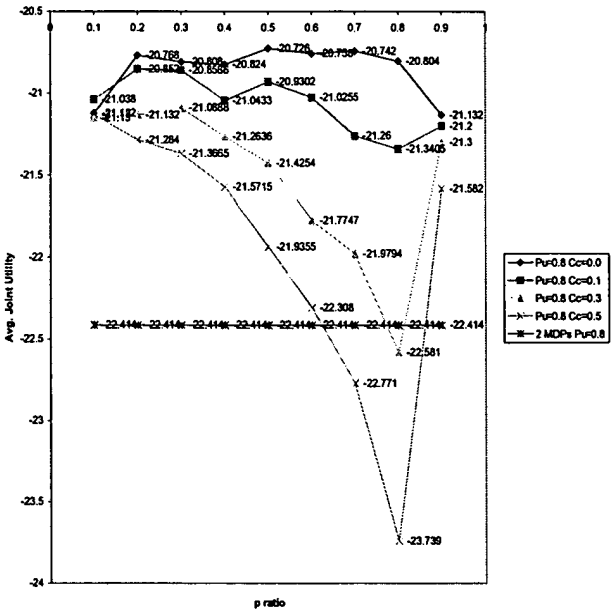

Figure 2: The average joint utility obtained when sub-goals are communicated.

In the Greedy case, we design the agents to optimize the time when a message will be sent assuming that they can communicate only once. At the off-line planning stage, the agents compute their expected joint cost, $\Theta_c(A, d, t)$, for every possible distance that can occur between the two agents, and for every time t (up to some maximal constant).

$\Theta_c(A, d, t)$ is the expected joint cost from taking control actions during t time steps, communicating at time $t + 1$ if the agents have not met so far, and following the optimal policy of control actions towards the expected goal without communicating (at an expected cost of $\Theta_{nc}(A, d_1, d_2)$ as computed for the No-Communication case). In the case that the agents met before the t time steps, then the ex-

pected cost considers the relevant expected joint cost that the agents incur until they met, i.e., less than t.
$\Theta_c(A,d,t) = \sum_{i=0}^{t}\sum_{j=0}^{t}\binom{t}{i}\binom{t}{j}P_u^i P_u^j(1-P_u)^{t-i}(1-P_u)^{t-j}\Phi(i,j)$

The function $\Phi(i,j)$ is the total cost that the agents incur after succeeding in moving i times for agent 1 and j times for agent 2. $\Phi(i,j) = 2R_a\tau(i,j)+\Theta_{nc}(A,d_1,d_2)+C_\Sigma Flag_c$. $\tau(i,j)$ is either t if the agents did not meet yet and otherwise we compute the expected number of time steps smaller than i and j when the agents succeeded to meet. d_1 and d_2 are updated based on the values of i and j.[8] $Flag_c$ is 1 if the agents have not met during the t time steps and therefore will incur a cost of C_Σ at time $t+1$, and 0 otherwise.

At each time t, each one of the agents know a meeting location, that is the last goal location that was synchronized. Each agent optimally moves towards this goal. In addition the optimal policy for single-communication is found by computing the earliest time t, for which $\Theta_c(A,d,t) < \Theta_{nc}(A,d_1,d_2)$, that is what is the best time to communicate such that the expected cost is the least. The optimal policy of communication is a table where each row specifies a time t when to communicate given a known distance between the agents.

We found the optimal single-communication policies for agents solving the meeting under uncertainty problem given that P_u takes the following values: 0.2, 0.4, 0.6, 0.8, the cost of taking a control action is $R_a = -1.0$ and the cost of communicating $C_\Sigma = -0.1, -1.0, -10.0$. For the smallest cost tested, it is always beneficial to communicate rather early, no matter the uncertainty in the environment, and almost no matter what is d_0 (the differences in time are between 2 and 4). For larger costs of communication for a given P_u, agents will communicate later as long as their distance is larger (e.g., when $P_u = 0.4, C_\Sigma = -1$ and $d = 5$, agents should communicate at time 4, but if $C_\Sigma = -10$, they should communicate at time 9). For a given C_Σ as long as the distance is larger the agents will communicate later (e.g., when $P_u = 0.4, C_\Sigma = -10$ and $d = 5$, agents should communicate at time 9, but if $d = 12$, they should communicate at time 16). The results from averaging over 1000 runs show that for a given cost C_Σ as long as P_u decreases (the agent is more uncertain about its actions' outcomes), the agents communicate more times.

In the 1000 experiments run, agents start at an initial state which is synchronized and therefore d_0 is known to both agents. The agents exchange information about their actual locations at the best time that was myopically found for d_0. After they communicate, they know the actual distance between them, denote it by d_t. The agents follow the same optimal single-communication policy to find the next time when they should communicate if they did not meet. This time is the best time that was found by the greedy algorithm given that the distance between the agents was d_t. Iteratively, the agents approximate the optimal solution to the decentralized control problem with communication by following their independent optimal policies of action, and the myopic policy for single-communication. Results obtained from averaging the joint utility attained after 1000 experiments show that these greedy agents perform better than agents who communicate sub-goals (that is a more efficient approach than no communicating at all). The results for $C_\Sigma = -0.1$ are presented in Tables 1 and 2.

[8]Due to space limits we do not include the details here.

	Average Joint Utility			
P_u	No-Comm.	Ideal	SubGoals[9]	Greedy
0.2	-104.925	-62.872	-64.7399	-63.76
0.4	-51.4522	-37.33	-38.172	-37.338
0.6	-33.4955	-26.444	-27.232	-26.666
0.8	-24.3202	-20.584	-20.852	-20.704

Table 1: $C_\Sigma = -0.10, R_a = -1.0$.

The Greedy approach attained utilities significantly greater than those obtained by the heuristic case when $C_\Sigma = -0.1$. Ideal always attained higher utilities than Greedy, but when $C_\Sigma = -0.1$ and $P_u = 0.4$ both values were not significantly different with probability 98%. When $C_\Sigma = -1$ the utilities attained for the Greedy approach when $P_u < 0.8$ are significantly greater than the results obtained in the heuristic case and for $P_u = 0.8$, the heuristic case for the best p was found to be better than Greedy (Greedy obtained -21.3, and the Subgoals with $p = 0.1$ attained -21.05 (variance=2.18)). The utilities attained by the Greedy agents, when $C_\Sigma = -10$ and $P_u = 0.2, 0.4$, were not significantly different than the SubGoals case for the best p with probabilities 61% and 82%, respectively. However, the heuristic case yielded smaller costs for the other values of $P_u = 0.6, 0.8$. One important point to notice is that these results consider the best p found for the heuristic, in general if a designer does not know this value then all the utilities obtained by Greedy were higher than the utilities attained for the worst p in the SubGoals case.

	Average Communication Acts Performed			
P_u	No-Comm.	Ideal $C_\Sigma = 0$	SubGoals	Greedy
0.2	0	31.436	5.4	21.096
0.4	0	18.665	1	11.962
0.6	0	13.426	1	8.323
0.8	0	10.292	1	4.579

Table 2: $C_\Sigma = -0.10, R_a = -1.0$.

For the same parameters tested so far, experiments were run with two deadlines, $T = 8, 15$. In general, the greedy policy found by a myopic agent may instruct the agent not to communicate if $\Theta_{nc} < \Theta_c$, i.e., had the agents communicated, unnecessary information had been exchanged. On the other hand, this policy may instruct an agent not to communicate, if given a deadline, the agent is not going to be able to reach the goal. In the first case, limiting the deadline to be earlier, results in policies of communication that stipulate that the agent should communicate earlier than in the case when no deadlines are added (for large values of d_0 with low uncertainties P_u). When no deadlines are assumed, the agents may benefit from exchanging information later. When a short deadline is assumed, if the agents have the chance to meet without communication given a later deadline, they will need to communicate earlier if the time stipulated in the policy with no deadlines is larger than the deadline. If the deadline is large enough for these agents to meet, they do not need to communicate at all. For shorter d values if the policy with no deadline allows the agent to communicate at a time smaller than the deadline the same policy holds.

[9]The results are presented for the best p, found empirically.

In the second case, the agents may not communicate if they may not meet at all by the stipulated deadline. The empirical results show that by extending the deadline, agents benefit from communicating at a time that is later than the time found by the myopic policy when no deadlines were assumed. Since, there is a chance of not meeting at all, agents need to wait until it becomes beneficial to communicate.

4. CONCLUSIONS

We have developed a theoretical formal model for decentralized control with communication extending current models based on Markov Decision Processes. This model enables the study of the trade-off between the cost of information and the value of the information acquired in the communication process and its influence on the joint utility of the agents. We have also discussed aspects relevant to the study of decentralized control with communication, such as languages of communication and cost models of communication. The analysis raises interesting questions regarding the design of communication languages, their semantics and their impact on the complexity of coordination. These open questions remain the focus of further study.

This paper represents one of the few formal studies of the hard problem deriving both actions and communication policies for decentralized cooperative multi-agent systems. We develop a precise formal definition of the value function and the optimization problem. A practical contribution of the paper is the development of a greedy meta-level approach to communication. This paper evaluates the greedy approach and shows that it produces near-optimal solutions in a simple testbed. Although a heuristic approach may perform well, it requires the correct tuning of its parameters in contrast to the greedy approach which requires none. Moreover, the greedy algorithm always outperforms the heuristic with the worst setting. Given the complexity of the problem of computing communication policies in general, we believe this is a promising direction to pursue in more complex domains. We are currently testing additional scenarios and aspects of the problem presented in this paper, including more complex models of deadlines, asymmetric uncertainties, and messages with partial information.

5. ACKNOWLEDGMENTS

This work was supported in part by NSF under grant IIS-9907331, by AFOSR under grant F49620-03-1-0090 and by NASA under grant NCC-2-1311. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of the NSF, AFOSR or NASA.

6. REFERENCES

[1] D. Bernstein, R. Givan, N. Immerman, and S. Zilberstein. The complexity of decentralized control of Markov decision processes. *Mathematics of Operations Research*, 27(4):819—840, 2002.

[2] C. Boutilier. Sequential optimality and coordination in multiagent systems. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, pages 478–485, Stockholm, Sweden, 1999.

[3] C. Claus and C. Boutilier. The dynamics of reinforcement learning in cooperative multiagent systems. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence*, pages 746–752, Madison, WI, 1998.

[4] K. S. Decker and V. R. Lesser. Generalizing the partial global planning algorithm. *International Journal of Intelligent Cooperative Information Systems*, 1(2):319–346, 1992.

[5] E. H. Durfee. *Coordination of Distributed Problem Solvers*. Kluwer Academic Publishers, Boston, 1988.

[6] P. J. Gmytrasiewicz, M. Summers, and D. Gopal. Toward automated evolution of agent communication languages. In *Proceedings of the 35th Hawaii International Conference on System Sciences*, 2002.

[7] D. Goldberg and M. J. Mataric. Coordinating mobile robot group behavior using a model of interaction dynamics. In *Proceedings of the Third International Conference on Autonomous Agents*, pages 100–107, Seattle, Washington, 1999.

[8] C. V. Goldman and J. S. Rosenschein. Emergent coordination through the use of cooperative state-changing rules. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, pages 408–413, Seattle, WA, 1994.

[9] J. Grass and S. Zilberstein. A value-driven system for autonomous information gathering. *Journal of Intelligent Information Systems*, 14:5–27, 2000.

[10] B. J. Grosz and S. Kraus. Collaborative plans for complex group action. *Artificial Intelligence*, 86(2):269—357, 1996.

[11] J. Hu and M. P. Wellman. Multiagent reinforcement learning: Theoretical framework and an algorithm. In *Proceedings of the the Fifteenth International Conference on Machine Learning*, pages 242–250, Madison, WI, 1998.

[12] M. L. Littman. Markov games as a framework for multi-agent reinforcement learning. In *Proceedings of the Eleventh International Conference on Machine Learning*, pages 157—163, 1994.

[13] M. J. Mataric. Learning social behaviors. *Robotics and Autonomous Systems*, 20:191–204, 1997.

[14] L. Peshkin, K.-E. Kim, N. Meuleau, and L. P. Kaelbling. Learning to cooperate via policy search. In *Proceedings of the the Sixteenth Conference on Uncertainty in Artificial Intelligence*, pages 489–496, Stanford, CA, 2000.

[15] D. V. Pynadath and M. Tambe. The communicative multiagent team decision problem: Analyzing teamwork theories and models. *Journal of Artificial Intelligence Research*, 16:389–423, 2002.

[16] S. Russell and E. Wefald. Principles of metareasoning. *Artificial Intelligence*, 49:361–395, 1991.

[17] J. Wang and L. Gasser. Mutual online concept learning for multiple agents. In *Proceedings of the First International Joint Conference on Autonomous Agents and Multi-Agent Systems*, pages 362–369, Bologna, Italy, 2002.

[18] P. Xuan, V. Lesser, and S. Zilberstein. Communication decisions in multi-agent cooperation: Model and experiments. In *Proceedings of the Fifth International Conference on Autonomous Agents*, pages 616–623, Montreal, Canada, 2001.